LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models

Year: 2023 | *Citations: 215* | *Authors: Yukang Chen, Shengju Qian, Haotian Tang*

Abstract

We present LongLoRA, an efficient fine-tuning approach that extends the context sizes of pre-trained large language models (LLMs), with limited computation cost. Typically, training LLMs with long context sizes is computationally expensive, requiring extensive training hours and GPU resources. For example, training on the context length of 8192 needs 16x computational costs in self-attention layers as that of 2048. In this paper, we speed up the context extension of LLMs in two aspects. On the one hand, although dense global attention is needed during inference, fine-tuning the model can be effectively and efficiently done by sparse local attention. The proposed shifted sparse attention effectively enables context extension, leading to non-trivial computation saving with similar performance to fine-tuning with vanilla attention. Particularly, it can be implemented with only two lines of code in training, while being optional in inference. On the other hand, we revisit the parameter-efficient fine-tuning regime for context expansion. Notably, we find that LoRA for context extension works well under the premise of trainable embedding and normalization. LongLoRA combines this improved LoRA with S^2-Attn. LongLoRA demonstrates strong empirical results on various tasks on Llama2 models from 7B/13B to 70B. LongLoRA extends Llama2 7B from 4k context to 100k, or Llama2 70B to 32k on a single 8x A100 machine. LongLoRA extends models' context while retaining their original architectures, and is compatible with most existing techniques, like Flash-Attention2. In addition, we further conduct supervised fine-tuning with LongLoRA and our long instruction-following LongAlpaca dataset.